Psychon Bull Rev (2014) 21:301–308
DOI 10.3758/s13423-014-0595-4

Optional stopping: No problem for Bayesians

Jeffrey N. Rouder

Published online: 22 March 2014

Abstract Optional stopping refers to the practice of peeking at data and then, based on the results, deciding whether or not to continue an experiment. In the context of ordinary significance-testing analysis, optional stopping is discouraged, because it necessarily leads to increased type I error rates over nominal values. This article addresses whether optional stopping is problematic for Bayesian inference with Bayes factors. Statisticians who developed Bayesian methods thought not, but this wisdom has been challenged by recent simulation results of Yu, Sprenger, Thomas, and Dougherty (2013) and Sanborn and Hills (2013). In this article, I show through simulation that the interpretation of Bayesian quantities does not depend on the stopping rule. Researchers using Bayesian methods may employ optional stopping in their own research and may provide Bayesian analysis of secondary data regardless of the employed stopping rule. I emphasize here the proper interpretation of Bayesian quantities as measures of subjective belief on theoretical positions, the difference between frequentist and Bayesian interpretations, and the difficulty of using frequentist intuition to conceptualize the Bayesian approach.

Keywords Optional stopping · Bayesian testing · p-hacking · Statistics · Bayes factors

The field of psychology is experiencing a crisis of confidence, as many researchers believe published results are not as well supported as claimed (Carpenter, 2012; Roediger, 2012; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Young, 2012). This crisis is comprised of publicized failures to replicate claimed effects, the publication of fantastic ESP claims, and the documentation of outright fraud. A common focus is now on identifying practices that violate the assumptions of our methods, and examples include peeking at the results to decide whether to collect more data (called *optional stopping*) and making additional inferential comparisons that were not considered before data collection. These questionable practices go under the moniker of *p-hacking*, and a remedy for the crisis is to avoid these bad practices (Simmons, Nelson, & Simonsohn, 2011).

An alternative viewpoint about the cause of the crisis is that the dominant inferential method, significance testing, is inappropriate for scientific reasoning (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2014). Many who are critical of significance testing recommend inference by Bayes factor as a replacement (Edwards, Lindman, & Savage, 1963; Gallistel, 2009; Myung & Pitt, 1997; Rouder et al., 2014; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Sprenger et al., 2013; Wagenmakers, 2007). The Bayes factor comes from Bayesian analysis and results from using Bayes's rule to update beliefs about theoretical positions after observing experimental data.

This article is about the wisdom of optional stopping, where the researcher collects some data, analyzes them, and on the basis of the outcome, decides to proceed with more data collection or not. Optional stopping is considered one of those bad p-hacking practices because it does affect conclusions from conventional significance tests. Yu, Sprenger, Thomas, and Dougherty (2013) have shown that common practices inflate both type I and type II error rates. Despite these results, there is a sense in which optional stopping seems like a smart thing to do. We seemingly should monitor our results as they come in, and we should end early when the results are clear and perhaps keep going when they are not. The critical question addressed here is whether optional stopping is problematic in the Bayesian context.

The answer to this question seems like it should be straightforward, yet the literature is contradictory. On one hand, early Bayesian theorists stated that Bayesian quantities are

J. N. Rouder (✉)
Department of Psychological Sciences, University of Missouri, 210 McAlester Hall, Columbia, MO 65211, USA
e-mail: rouderj@missouri.edu

interpretable under optional stopping. Lindley (1957) wrote, "It follows that any significance test based on Bayes's theorem does not depend on the sequential stopping rule used, at least amongst a wide class of such rules. In the extreme case the experimenter can go on sampling until he [has a Bayes factor that] has reached the significance level *c*, and yet the fact that he did so is irrelevant to a Bayesian" (p. 192). Likewise, Edwards et al. (1963) wrote, "the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience" (p. 193). Based on this earlier work, Wagenmakers and colleagues now use and recommend optional stopping in experimental design (Matzke et al., 2014; Wagenmakers et al., 2012).

More recently, Yu et al. (2013) have called this sanguine answer into question. They write,

> Bayesian analysis is not the magic elixir it is sometimes made out to be. One cannot simply apply Bayesian statistics to any old dataset and be confident that the outcome is free of bias. As our results illustrate, the BF [Bayes factor] distribution shows substantial irregularities, which vary depending on which heuristic was used to collect data. Thus, prior analysis in which BFs are computed post hoc on data collected under the NHST framework . . . are not interpretable if researchers used a data-dependent stopping heuristic. (p. 32)

According to Yu et al., Bayesian methods are susceptible to an optional-stopping-rule artifact, and researchers cannot reanalyze previously collected data unless they are certain that the original researcher did not use an optional-stopping rule. Such a ramification would certainly put me and my colleagues at a loss, as we have reanalyzed others' data (see Rouder & Morey, 2011; Rouder, Morey, & Province, 2013), yet we remain unsure whether these studies were terminated optionally or not. Sanborn and Hills (2013) present a far more nuanced argument, but their conclusions are not so different. These authors write that under a reasonable interpretation of the Bayes factor, "the choice of stopping rule can, in some situations, greatly increase the chance of an experimenter finding evidence in the direction they desire" (p. 1).

This article provides a critique of Yu et al. (2013) and Sanborn and Hills (2013). Before proceeding, please note that both Yu et al. and Sanborn and Hills make a number of important contributions, and both papers recommend the Bayes factor for inference, at least under certain circumstances. Moreover, Yu et al.'s main contribution was documenting the degree of optional-stopping tendencies among practicing researchers, and their critique of Bayesian methods was secondary. The critical element addressed here is whether optional stopping is problematic for Bayesians. My argument is that both sets of authors use the wrong criteria or lens to draw their conclusions. They evaluate and interpret Bayesian statistics as if they were frequentist statistics. The more germane question is whether Bayesian statistics are interpretable *as Bayesian statistics* even if data are collected under optional stopping.

Bayesian probability and model comparison

Most of us were taught the *frequentist* definition of probability: Probability is a proportion in the long run. For example, the probability that a flipped coin lands heads is the proportion of heads in a very long series of flips. Frequentist probability has substantial limits. For example, it cannot be used on events that occur only once—say, the probability that the Euro will collapse in the next decade—because there is no concept of a long-run series (Jackman, 2009; cf. Hájek, 2007). Likewise, because there is no concept of a long-run series, probabilities may not be placed on models, hypotheses, or theories. Bayesian analysts use probability to express a degree of belief. For a flipped coin, a probability of 3/4 means that the analyst believes it is three times more likely that the coin will land heads than tails. Such a conceptualization is very convenient in science, where researchers hold beliefs about the plausibility of theories, hypotheses, and models that may be updated as new data become available. Not only does Bayesian probability quantify these beliefs, but also Bayes Rule provides the ideal way of updating these beliefs as new data become available. Bayes rule comes from careful consideration of what plausibility means, and its logical foundations may be found in Cox (1946), de Finetti (1995), Ramsey (1931), and Savage (1972).

Here is how significance testing and Bayesian model comparison work: Let's suppose that we wish to test the proposition that I can change the probability a coin lands tails simply by asking the coin to do so. Let's consider the null model that the coins true probability of a tails is .5 versus an effect-model that the true probability is .525, which is a 5% effect. Let's further suppose that I have asked 1,000 coins to land tails, and after flipping each a single time, 527 of them do so. Figure 1 shows the probability of all outcomes under the null model (open and light points) and under the 5% effect model (filled points). First, let's consider a significance test. To perform a significance test, we calibrate our assessment to the null model alone. As can be seen, 527, the value at the vertical line, is a rare event under the null. It is so rare that the probability of observing it or any greater number of tails (light points) is less than 5%. Hence, we may reject the null at the conventional level of $p<.05$, and I have now documented my coin cajoling skills.

In the Bayesian approach, we may place probabilities directly on the models and then update these probabilities in

light of the data. We start with our beliefs before seeing the data. Whereas my coin-cajoling skills are at stake, I might be inclined to believe that I am as likely as not to affect the probability. You, of course, may be more skeptical and may hold odds, say, of a million-to-one against my purported coin-cajoling skills. This formulation of beliefs as odds is very convenient and is retained throughout. Next, we update our odds in light of the data, using Bayes rule:

$$\frac{P\left(M_1\middle|\text{Data}\right)}{P\left(M_0\middle|\text{Data}\right)}=\frac{P\left(\text{Data}\middle|M_1\right)}{P\left(\text{Data}\middle|M_0\right)}\times\frac{P(M_1)}{P(M_0)}, \quad (1)$$

where M_1 and M_0 denote the effect model and null model, respectively. The left-hand term, $P(M_1|\text{Data})/P(M_0|\text{Data})$ is the posterior odds, the relative beliefs about the models after seeing the data. The rightmost term, $P(M_1)/P(M_0)$ is the prior odds, the relative beliefs before seeing the data. The term $P(\text{Data}|M_1)/P(\text{Data}|M_0)$ is the *Bayes factor*, and it describes how beliefs are to be updated. Evaluating the Bayes factor is straightforward for the coin-cajoling example. We can use Fig. 1 for the values. At the observed data of 527 tails, the probability of this result under the alternative is .025, and the probability under the null is .0058. The ratio, the Bayes factor, is 4.3. I may now update my posterior odds to 4.3-to-1 in favor of the existence of my abilities; you may update yours to approximately 235,000-to-1 against. I am now modestly positive about my abilities; you are slightly less skeptical. Note that even though we may not share posterior beliefs, we can agree on how the data should obligate us to update our beliefs. Bayes factors serves as an appropriate and transparent measure of evidence from data for theoretical positions, and we, along with many others, recommend it be reported, rather than significance test results.

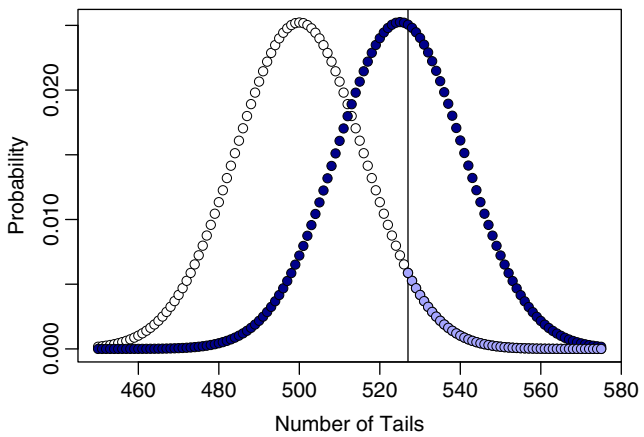

Fig. 1 The probability of a certain number of tail-side flips out of 1000 for $p = .5$ (open and light points) and for $p = .525$ (dark points). A value of 527 is significant by a one-tail test at .05. Nonetheless, the value is only 4.3 times as probable under the alternative as under the null

The interpretation of posterior odds holds with optional stopping

The main question here is whether Bayesian analysis is interpretable even with optional stopping. Yu et al. (2013) and Sanborn and Hills (2013) use computer simulations, rather than mathematical derivation, to elucidate the properties of analytic methods. This choice is wise for a readership of experimental psychologists. Simulation results have a tangible, experimental feel; moreover, if something is true mathematically, we should be able to see it in simulation as well.

Suppose a researcher is considering two hypotheses: a null with an effect size of 0 and an alternative hypothesis with an effect size of .4. Now let's generate some data—say, 10 observations from one of the hypotheses.[1] Moreover, let's pick which hypothesis we use to generate the data by flipping a fair coin. Before observing the data, let's set our prior odds to 1-to-1; after all, the generating hypothesis is chosen by coin flip. Now, let's observe the data and update our beliefs. The updated beliefs are the posteriori odds, or the probability that the data came from the alternative, relative to the probability that the data came from the null, conditional on the data.[2] It may seem natural to study the distribution of the posterior odds when the data come from one or the other hypothesis, and Fig. 2a shows the distribution of posterior odds when the null is true. The distribution was constructed by simulation, and there are 20,000 replicate experiments of 10 samples each. As can be seen, most of the posterior odds across these repeated experiments favor the null; that is, they are smaller than 1.0 in value. Figure 2b shows the distribution when the alternative is true—that is, when the true effect size is .4. Although the results are reasonable and the distributions are well behaved, they do not address the interpretability of posterior odds.

Posterior odds are the probability of competing hypotheses given data. If updating through Bayes factor is ideal and if the prior odds are accurate, then the posterior odds should be accurate as well. If a replicate experiment yielded a posterior odds of 3.5-to-1 in favor of the null, then we expect that the null was 3.5 times as probable as the alternative to have produced the data. We can check this interpretation with simulations as follows: In repeated simulations, we can select

[1] Throughout this report, the standard deviation of data is assumed known, for simplicity. All results about the interpretation of the Bayes factor and posterior odds hold without this assumption.

[2] With 1-to-1 prior odds, the posterior odds are $\frac{P(M_1|\text{Data})}{P(M_0|\text{Data})} = \exp(n\delta[\bar{y}-\delta=2])$, where δ is the effect size under the alternative, n is the sample size, and $\bar{y}$ is the sample mean.

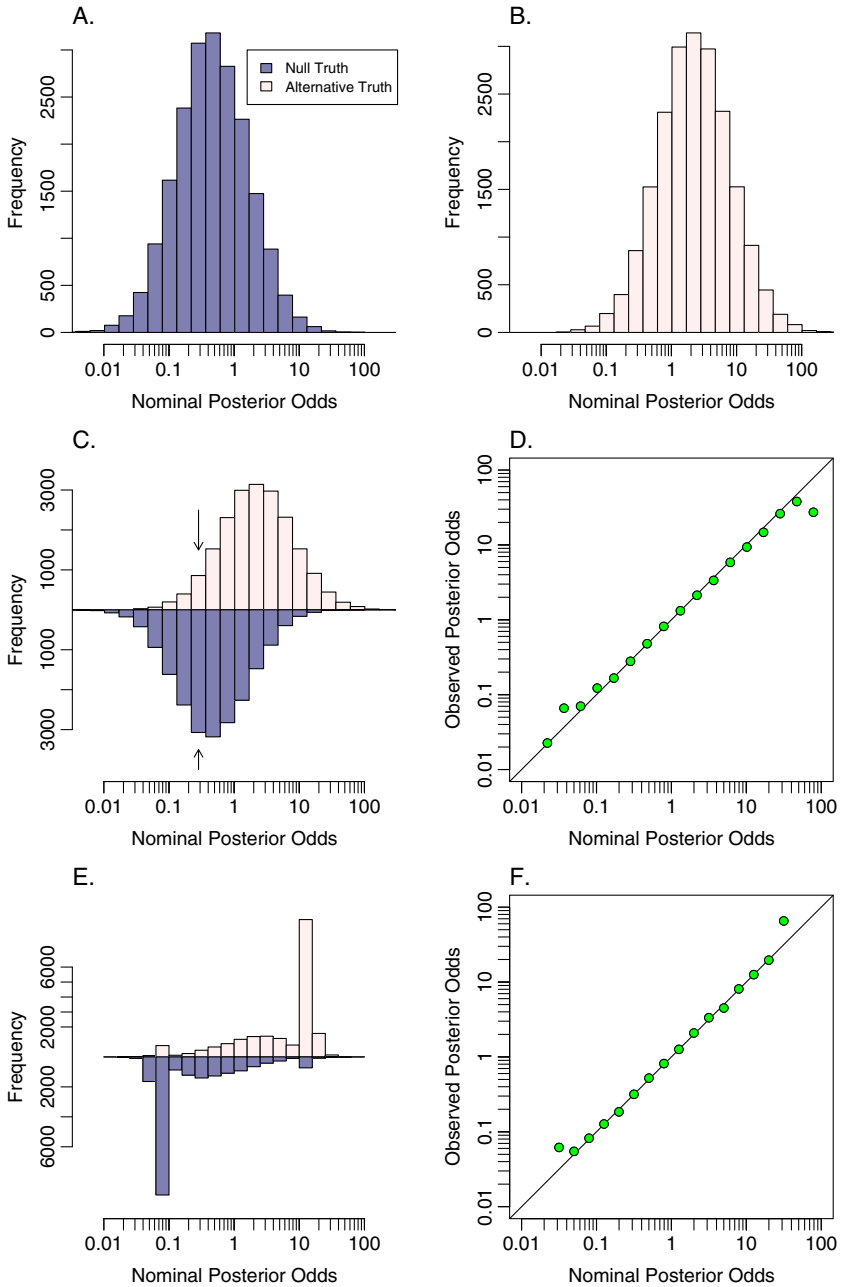

Fig. 2 The interpretation of posterior odds. **a** The distribution of posterior odds ($N = 10$, with 20,000 replicate experiments) under the null. **b** The same distribution under the alternative with an effect size of $\delta = .4$. **c** These distributions are displayed back-to-back, with the distribution for the null projected downward. This display allows for the selection of replicate experiments with similar posterior odds, regardless of the effect size that generated the data. The ratio of frequencies of replicates from each hypothesis that generated the data is the *observed posterior odds*. **d** Observed posterior odds as a function of nominal posterior odds. The equality holds within sampling noise. **e** Observed posterior odds with optional stopping. **f** Observed posterior odds as a function of nominal posterior odds with optional stopping

all those replicate experiments that yield the same posterior odds—say, 3.5-to-1 in favor of the null—and tally how many of these selected experiments came from the null truth and how many came from the alternative truth. If the posterior odds are interpretable as claimed, then about 3.5 times as many of these selected experiments should come from the null than from the alternative. Figure 2c shows the comparison. The histogram for the posterior odds from the alternative is shown as in Fig. 2b, but the histogram from the null is that from Fig. 2a projected downward. The arrow highlights a small bin of posterior odds centered on .284, which is about 3.5-to-1 in favor of the null. If the posterior odds are interpretable, there should be 3.5 times as many experiments when the null serves as truth (projected downward) as when the alternative serves as truth. In fact, for the 20,000 runs for each hypothesis, there were 3,072 and 858 runs in this small bin for when the null and alternative, respectively, served as truth. The ratio here, which for the purposes of this demonstration is called the *observed posterior odds*, is 3,072-to-858, which reduces to 3.6-to-1. This observed value agrees closely with the nominal value of 3.5-to-1, and the difference is well within the error of the simulation. Figure 2d shows the observed posterior odds as a function of the nominal posterior odds for all bins, and they are equal in value up to simulation

error (the small mismatches at the extreme values reflect small numbers under one or the other truths).[3]

Does the interpretation hold with optional stopping? I ran a simulation with the same setup, except that sampling occurred until the posterior odds were at least 10-to-1 for either hypothesis, unless a maximum of 25 samples was reached. With these settings, about 58% of the trials achieve the 10-to-1 criterion. The histograms of posterior odds under both hypotheses are shown in Fig. 2e, and once again, the posterior odds distribution for when the null served as truth is projected downward. These distributions may be compared with those from the previous simulation without optional stopping (shown in Fig. 2c), and they are quite different. The distributions no longer have a characteristic normal shape and, instead, have clumps at the stopping criteria of 10-to-1. Yu et al. (2013) describes these distributions as "irregular" and "distorted," in so much as they do not resemble those without optional stopping, and it is this feature that drives their conclusion. For Sanborn and Hills (2013), optional stopping is problematic when it changes the likelihood of obtaining posterior odds of certain values, which it does here. Yet these concerns are immaterial for the proper interpretation of posterior odds. The critical question is whether the posterior odds accurately reflect the probability that a given value came from a given hypothesis. Figure 2e shows the observed posterior odds as a function of nominal posterior odds for all bins, and as can be seen, they match up to simulation error. Optional stopping does not affect the interpretation of posterior odds. Even with optional stopping, a researcher can interpret the posterior odds as updated beliefs about hypotheses in light of data.

In Bayesian analysis, researchers can hold beliefs on hypotheses that encompass more than a single point, called *composite hypotheses*. Consider a model of the alternative in which the effect size is distributed as a standard normal. This alternative captures the belief that effect sizes as large as 1.0 are neither typical nor exceedingly rare. The posterior odds in this case are the updated beliefs about this normally distributed alternative, relative to the null.[4] I ran the above simulation where, in one case, the null served as truth and, in the other, the composite alternative served as truth. When the composite serves as truth, there is no single truth for all experiments. Instead, each experiment has a unique true value, yet these values are from a common distribution (the standard normal, in this case). In simulation, the true effect size is sampled for the replicate experiment, and then the n samples are generated from that true effect size. Following data generation, the posterior odds are then computed. In the fixed stopping case, the sample size was $n = 10$, and 20,000 replicate experiments were sampled. Figure 3a shows the distribution of posterior odds under this composite alternative (projected upward) and under the null (projected downward). Figure 3b shows that the observed posterior odds match the nominal values within simulation error.

The effects of optional stopping on the interpretation of posterior odds may be assessed as before. I performed the previous optional stopping simulation for the composite hypothesis case; sampling continued either until the posterior odds first exceeded 10-to-1 in favor of either hypothesis or until 25 samples were obtained. Figure 3c shows the histograms of posterior odds for both truths for this stopping rule. These distributions are different from those without optional stopping, and such differences are expected and unimportant in the interpretation of posterior odds. The critical question is whether the nominal and observed posterior odds match. Figure 3d shows that they do within simulation error. Hence, posterior odds are interpretable for composite hypotheses even with optional stopping.

I ran two additional simulations in which stopping was based on p-values: In one simulation, sampling continued until the p-value was less than .05 or until 25 samples were obtained. In the second simulation, sampling continued until the p-value was outside the interval from .05 to .50 or until 25 samples were obtained. This second simulation reflects the findings of Yu et al. (2013), who report that researchers also stopped when p-values became large. Figure 4 shows the results: The top row is for optional stopping based on small p-values; the bottom row is for optional stopping based on small and large p-values. Optional stopping based on p-values does not affect the interpretation of posterior odds. As an aside, the simulations show the inflation of type I error in significant tests. Consider the downward histogram in Fig. 3a, which is the distribution of posterior odds under the null. The second mode at about 2.5-to-1 in favor of the alternative consists of those replicate experiments that were terminated by optional stopping at $p < .05$. In fact, about 11.5% of the replicates met this criterion, meaning that although the nominal type I error rate was set at .05, the real type I error rate was .115.

Suppose the models are wrong?

In the above demonstrations, the analysts computed posterior odds for models that were used to generate the data. In real applications, however, there is no such guarantee that either model is true. One of Sanborn and Hills's (2013) demonstrations is to show that optional stopping affects the distribution of Bayes factors when the data come from a model not under

[3] The situation here is ironic in that the simulations provide estimates of odds from long-run frequencies. Nonetheless, the simulations are being used as a computational tool to assess whether updating is rational, and this usage is appropriate.

[4] With 1-to-1 prior odds, the posterior odds for the composite that effect sizes follow a standard normal are $\frac{P(M_1|\text{Data})}{P(M_0|\text{Data})} = \frac{\exp\left(\frac{n^2\bar{y}^2}{2(n+1)}\right)}{\sqrt{n+1}}$.

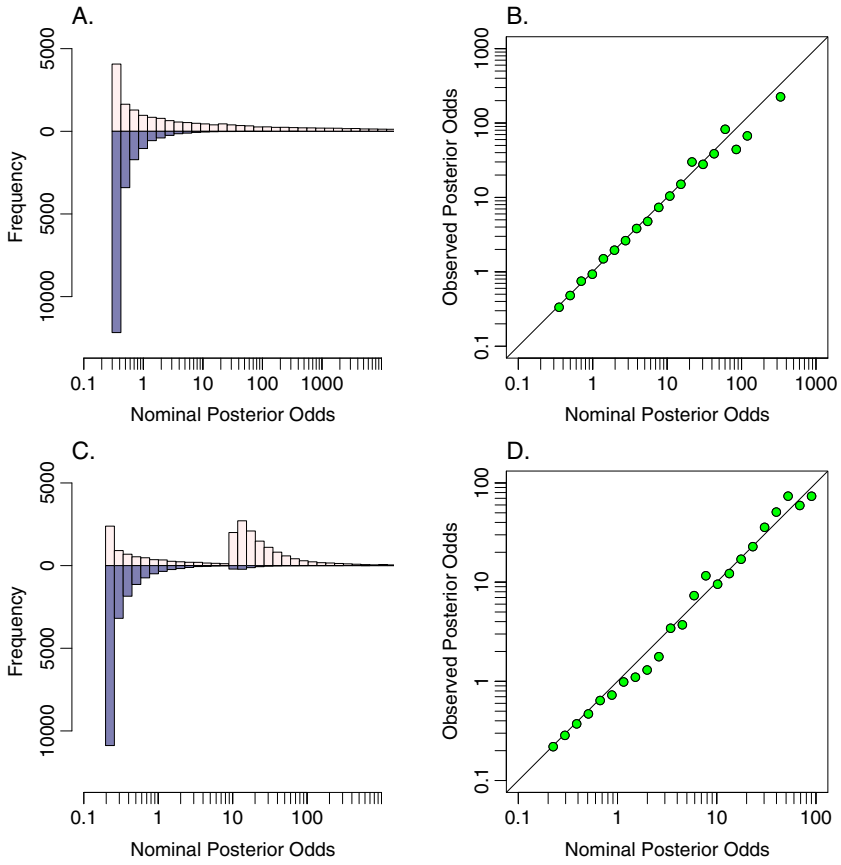

Fig. 3 Posterior odds are interpretable for composite hypotheses and with optional stopping. **a** The top histogram is posterior odds when the truth is distributed as a standard normal. The bottom histogram is the posterior odds under the null. **b** Observed posterior odds as a function of nominal posterior odds. The equality holds within sampling noise. **c** Observed posterior odds with optional stopping at 10-to-1 odds. **d** Observed posterior odds as a function of nominal posterior odds with optional stopping

consideration. Given the plausibility of some level of misspecification, it is wise to explore the interpretation of posterior odds for wrong models.

Bayesian analysts update beliefs about competing models, and, fortunately, updating does not require that any one model be true. The enterprise is motivated as follows: Bayesian analysts build models to capture differences between theoretically important positions. Beliefs may be placed on the models as surrogates for the positions, and these beliefs may be updated as new data are acquired. The resulting updated beliefs may be *interpreted* as the relative plausibility of these positions, at least inasmuch as the models captured the important relations between the positions (Morey, Romeijn, & Rouder, 2013). Such a view places a responsibility on the analyst to choose judicious models that indeed capture the relations between positions. Analysts benefit when they ask themselves what may be learned if the models are wrong, and it is often the case that very little may be learned if all models under consideration are dramatically misspecified. How to choose these models is a matter of some debate, but there is much agreement about the value of default models for common cases (see Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012; Rouder et al., 2013; Rouder et al., 2009; Wagenmakers, 2007; Wetzels, Grasman, & Wagenmakers, 2012).

Sanborn and Hills (2013) offer an example in which models are dramatically wrong, and the following example is similar in spirit: Considered two point hypotheses that the effect size was small and positive ($\delta = .2$) versus it was small and negative ($\delta = -.2$). Figure 5a shows the distribution of the posterior odds for $n = 40$ when the data are generated with $\delta = 0$, a truth represented by neither model. This posterior-odds distribution is centered at about 1-to-1, which is expected. Figure 5b shows the case for optional stopping. Here, sampling occurred until the odds were 10-to-1 in favor of the positive effect up to 80 samples. Note that the probability of reaching a 10-to-1 odds in favor of the positive result is greatly increased (in this case, reaching a 10-to-1 odds is twice as likely as reaching a 1-to-10 odds).[5] It is this fact that leads Sanborn and Hills to conclude that optional stopping may increase the chances of a desired result. These results, however, do not impinge on the interpretability of posterior odds. When we update relative beliefs about two models, we make an implicit assumption that they are worthy of our consideration. Under this assumption, the beliefs may be updated regardless of the stopping rule. In this case, the models are dramatically wrong, so much so that the posterior odds

[5] If the analyst is willing to sample enough data, then she or he can always end up at a posterior odds of 10-to-1 for the positive result (Feller, 1968).

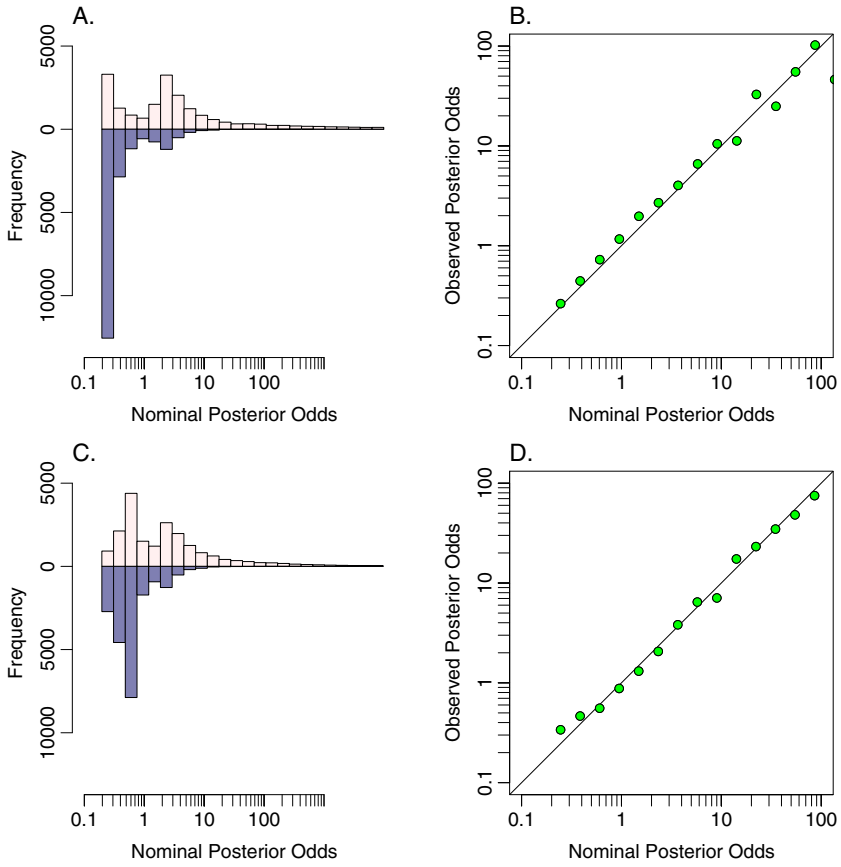

Fig. 4 Posterior odds are interpretable for composite truths and with optional stopping based on *p*-values. **a**–**b**. Posterior odds with optional stopping from small *p*-values ($p < .05$). **c**–**d** Posterior odds with option stopping when the *p*-value is outside the interval from .05 to .50

contain no useful information whatsoever. Perhaps the more important insight is not that optional stopping is undesirable, but that the meaningfulness of posterior odds is a function of the usefulness of the models being compared.

Discussion

As was discussed by early Bayesian theorists such as Lindley (1957) and Edwards et al. (1963), the proper interpretation of Bayesian statistics such as posterior odds and Bayes factors is unaffected by the stopping rule. The following three recommendations may prove helpful for substantive researchers.

1. Researchers should consider Bayesian methods to assess the evidence from data for theoretically important propositions, relative to judiciously chosen alternatives. Bayesian updating provides a rigorous and appealing approach to communicating results in scientific discourse. Enlightened advocacy may be found in a growing number

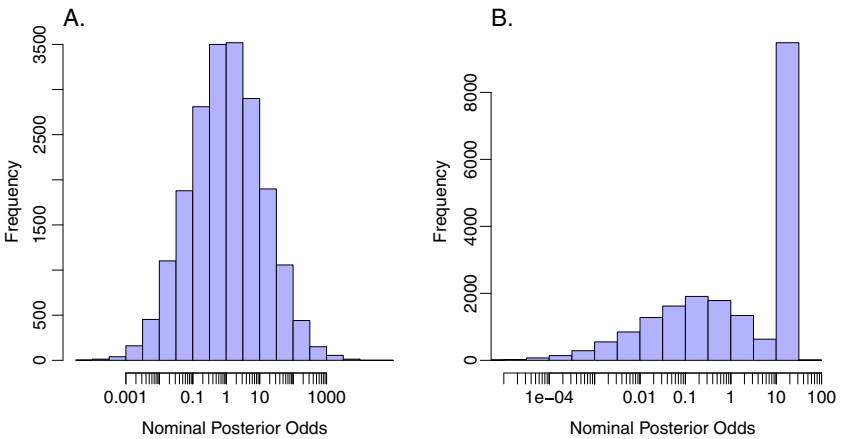

Fig. 5 Distributions of posterior odds for beliefs that the effect is small and positive versus it is small and negative when the null generates the data. **a** Distribution with fixed sample size of 40 samples. **b** Distribution with optional stopping for the positive hypothesis. The shape and form of these distributions are not indicative of the interpretability of Bayesian statistics

of sources, including Berger and Sellke (1987), Edwards et al. (1963), Jeffreys (1961), Rouder et al. (2009), and Wagenmakers (2007). How to choose models for common cases has been studied, and there is much gathered wisdom in the Bayesian psychology community.

2. Researchers who use Bayesian testing should use the proper interpretation as updated beliefs about the relative plausibility of models in light of data. The critical error of Yu et al. (2013) and Sanborn and Hills (2013) is studying Bayesian updating conditional on some hypothetical truth rather than conditional on data. This error is easy to make because it is what we have been taught and grown familiar with in our frequentist training. In my opinion, the key to understanding Bayesian analysis is to focus on the degree of belief for considered models, which need not and should not be calibrated relative to some hypothetical truth.
3. Bayesians should consider optional stopping in practice. Wagenmakers et al. (2012) recommended a protocol where researchers state, before data collection, that they will sample until the Bayes factor reaches sufficient size in favor of one model over the other, and Matzke et al. (2014) implemented this protocol. Optional-stopping protocols may be hybrids where sampling occurs until the Bayes factor reaches a certain level or a certain number of samples is reached. Such an approach strikes me as justifiable and reasonable, perhaps with the caveat that such protocols be made explicit before data collection. The benefit of this approach is that more resources may be devoted to more ambiguous experiments than to clear ones.

Acknowledgment This research was supported by National Science Foundation grants BCS-1240359 and SES-102408.

References

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association, 82*(397), 112–122. http://www.jstor.org/stable/2289131.

Carpenter, S. (2012). Psychology's bold initiative. *Science, 335,* 1558–1561.

Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American Journal of Physics, 14,* 1–13.

de Finetti, B. (1995). The logic of probability. *Philosophical Studies, 77,* 181–190.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70,* 193–242.

Feller, W. (1968). *Introduction to probability theory and its applications* (3rd ed., Vol. 1). New York: Wiley.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116,* 439–453. http://psycnet.apa.org/doi/10.1037/a0015251.

Hájek. (2007). The reference class problem is your problem too. *Synthese, 156,* 563–585.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester: John Wiley & Sons.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York: Oxford University Press.

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44,* 187–192.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E. J. (2014). Two birds with one stone: A preregistered adversarial collaboration on horizontal eye movements in free recall. Manuscript submitted for publication.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2013). The humble Bayesian: model checking from a fully Bayesian perspective. *British Journal of Mathematical and Statistical Psychology, 66,* 68–75.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin and Review, 4,* 79–95.

Ramsey, F. P. (1931). *The foundations of mathematics*. London: Routledge and Kegan Paul.

Roediger, H. L. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer, 25*.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18,* 682–689. http://dx.doi.org/10.3758/s13423-011-0088-7.

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research, 47,* 877–903.

Rouder, J. N., Morey, R. D., & Province, J. M. (2013). A Bayes-factor meta-analysis of recent ESP experiments: A rejoinder to Storm, Tressoldi, and Di Risio (2010). *Psychological Bulletin, 139*, 241-247.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56,* 356–374.

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2014). *The $p < .05$ rule and the hidden costs of the free lunch in inference.* Manuscript under review.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review, 16,* 225–237. http://dx.doi.org/10.3758/PBR.16.2.225.

Sanborn, A. N., & Hills, T. T. (2013). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*.

Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366.

Sprenger, A. M., Atkins, S. M., Bolger, D. J., Harbison, J. I., Novick, J. M., Chrabaszcz, J. S., & Dougherty, M. R. (2013). Training working memory: Limits of transfer. *Intelligence, 41*(5), 638–663. http://www.sciencedirect.com/science/article.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problem of *p* values. *Psychonomic Bulletin and Review, 14,* 779–804.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7,* 627–633.

Wetzels, R., Grasman, R. P., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for ANOVA designs. *American Statistician, 66,* 104–111.

Young, E. (2012). Nobel laureate challenges psychologists to clean up their act: Social–priming research needs "daisy chain" of replication. *Nature*.

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2013). When decision heuristics and science collide. *Psychonomic Bulletin & Review*.